Interactive visualization of high-dimensional data via continuous low-dimensional projections

Mid-candidature review — 25 Feburuary, 2020

Nicholas Spyrison, B.Sc

Monash University

Faculty of Information Technology

Thesis Supervisors

Prof. Kimbal Marriott

Prof. Dianne Cook

Committee Members

Dr. Maxime Cordiel

Dr. Shirui Pan

Chair

Assoc. Prof. Bernhard Jenny

Contents

1 Introduction

1.1 Motivation

The term exploratory data analysis was coined by Tukey (1977), who leaves it as an intentionally broad term that encompasses the initial summarization and visualization of a data set. This is a critical first step of checking for realistic values and validating model assumptions. It may be tempting to review a series of summary statistics to check model assumptions. However, there are known datasets where the same summary statistics miss glaringly obvious visual patterns (Anscombe 1973; Matejka and Fitzmaurice 2017). It is strikingly simple to look at the wrong, or incomplete set of statistics needed to validate assumptions. Data visualization is fast, versatile, and robust relative to the alternative of numeric statistical summarization. Data visualization does and must remain a primary component of data analysis and model validation.

Consider tabular data containing many attributes (I will use attribute and variable as synonymous). Visualization of this ubiquitous type of dataset is key to its understanding and exploration. Visualization of spaces in more than 3 dimensions quickly becomes problematic. We will discuss the use of linear projections to mitigate this obstacle. The motivation for this research is two-fold: expand the dimensionality support and improve the understanding from such visualizations.

1.2 Current state of the field

Consider plotting 2 variables as an XY scatterplot. To add a 3rd variable append the z-axis orthogonal (right angle to) the XY plane. Increasing dimensionality furth is not so trivial solved. To resolve this, we will introduce the basis vector and linear projections before traditional *discreate* method, more recent *continuous* methods and then discuss user control

Mathematically, a 'basis' is an orthonormal set of `d` element vectors, where each element vector is a linear combination of `p` elements in vector-space. That is to say, a basis defines the direction and magnitude each of the `p` attributes contributes to every direction of the `d`-dimensional projection space. Consider a projection down to 2D. The basis defining this projection can be visually displayed as a unit circle with `p` line segments stemming from the origin as shown in figure 1.

Many methods identify one or more projections of interest. For instance, scatterplot matrices (Chambers et al. 1983) can be used to rapidly valid univariate supports and distributions. Principal component analysis (PCA; Pearson 1901) and biplots (Gabriel 1971) highlight full sample variation. Linear discriminant analysis (Fisher 1936) and penalized discriminant analysis (Hastie, Buja, and Tibshirani 1995) can be used to identify

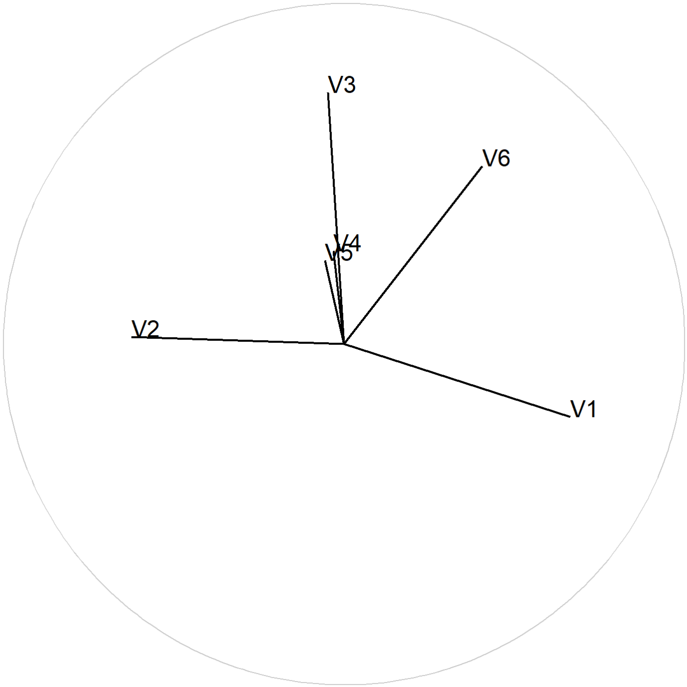


Figure 1: Illustration of a basis projecting 6 attributes down to 2 dimensions. The black line segments indicate the magnitude and direction a variable contributes to the projection.

projection with cluster separation. These are great techniques and all applicants for numeric data with many attributes, however, they have two notable features: they show only 1 or several *discrete* projections and they do not allow for an analyst to steer or control.

1.3 From discrete to continuous

The above methods have suggested a *discrete* number of linear projections to look at. At the same time, the stool example illustrates that looking at intermediate views improves understanding. A *continuous* animation of the object being rotated would improve this understanding even further. This is analogous to the idea of a data visualization *tour*(Asimov 1985; Buja and Asimov 1986). A tour produces a relatively high number of linear projections and views them in quick succession, typically as an animation. When the bases have a relatively small change in the contributions the projections are much closer. Single points and features can be tracked and follow from projection to projection allowing for a better understanding of the local structure.

The path that a tour animates is a crucial feature. There are various types of tours, many of which select an arbitrary or function-driven path. All of the discrete techniques also highlight a specific feature. For EDA it is desirable to give the analyst the ability to choose the direction to explore. To explore higher dimensional spaces visually, we need *human-in-the-loop* (Karwowski 2006) tools. (Cook and Buja 1997) introduces the idea of the *manual* tour. In a manual tour, an individual variable is selected and its contribution to the projection basis is then controlled. This allows analysts to choose the direction of exploration.

Figure 2 illustrates a manipulation space. the basis we viewed before is shown in blue and unselected variables in grey on a horizontal plane. while the manipulation space extends above it in red. The contribution that V1 has on the blue projection plane can be controlled as though a hand might manipulate the end of the red line segment. The other variables would perform a constrained rotation, maintain orthogonality of the space.

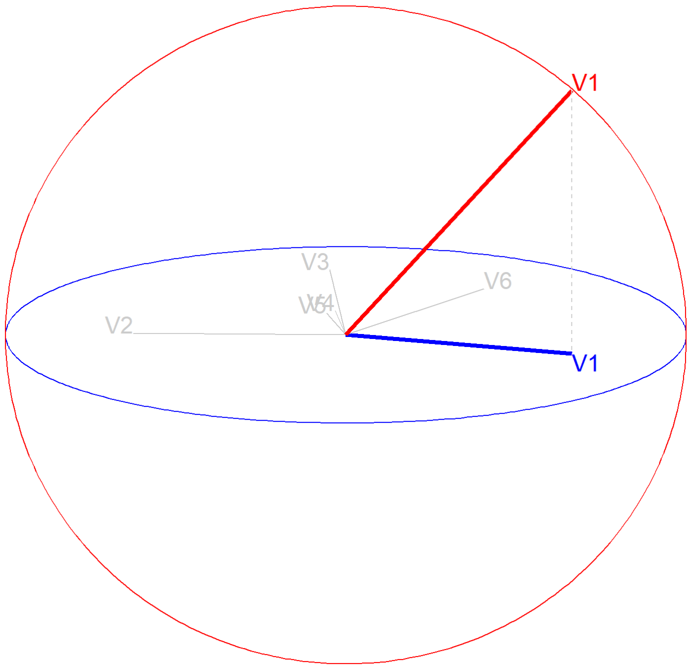


Figure 2: Visualization of a manipulation space. The projection plane extends horizontally in blue and grey. The manipulation direction in red extends orthogonal to this. Variable 1 is selected as the manipulation variable. Its contribution to the projection plane (blue) is controlled by its position in the manipulation space (red). If the red segment is manipulated all variable change due to orthogonally-constrained rotation.

The discrete methods above identity discrete bases that highlight some feature. The above tours can help convey more continuity of structure through many small changes in the basis. For EDA it is desirable to give the analyst the ability to choose the direction to explore.

1.4 Steering 2-D projections

The manual tour can play a predefined manual path or be used interactively, where the user defines each subsequent step while viewing the current projection. The later, human-in-the-loop method we define as

User-Controlled Steering (UCS) specifically as the interactive, human-in-the-loop application of the manual tour.

User-control manipulation around identified projections of interest theoretically allows for a better understanding of the variables contributing to the structure. However, the level of abstraction and the sheer number of basis permutations is formidable. The idea that UCS will provide digestible improvements to the understanding of structure should be validated.

1.5 Research objectives

Data and models are typically high-dimensional, with many variables and parameters. Developing new methods to visualize high dimensions has been a pursuit of statisticians, computer scientists and visualization researchers for decades. Their continues to be tension and back-and-forth on automated- and manual-analysis. Hardware advances allow for more realistic virtual-, augmented-, and mixed- environments with ever-increasing computing power. The degree of analyst choice and interaction with the display dimension should be explored. The primary research objectives (RO) can then be summarized as the following.

Overall objective: **Can a geodesic interpolator be used to help analysts understand linear projections of data, and explore the sensitivity of structure in the projection to the variables contributing to the projection?**

1. **How do we define a geodesic interpolator be used to add and remove variables smoothly from a 2D linear projection of data?**
   An algorithm for the manual tour is discussed in Cook and Buja (1997). The algorithm will be adapted to create the geodesic interpolator. Input will be a single projection, a full p-dimensional data matrix, and a choice of variable to control. The implementation will be in R.

2. **Do analysts understand the relationship between variables and structure in a 2D linear projection better when the geodesic interpolator is available?**
   Current practice for high-dimensional data is to make a single low-dimensional representation, with a method such as principal component analysis (PCA). Often just the first two principal components are shown, which is equivalent to a 2D projection. This technique is also called a biplot. Another more rarely used technique is a tour, which shows an animation of 2D projections. It can be used to get an overview of the structure in the multivariate space, but it is difficult to assess the importance of variables to a particularly interesting low-dimensional structure. A human subject study has been

designed and conducted to assess the learning about the importance of variables, from each of these techniques.

3. **Can we define a geodesic interpolator for 3D projections, so that the technology can be implemented in modern virtual reality environments?**
   The cutting edge of visualization research today is in virtual environments, as technology has advanced to make this accessible to the masses. It is interesting to explore how analysis with 3D graphics operates in comparison with 2D graphics. Building a geodesic interpolator for 3D projections extends the technology from 2D to 3D, and will allow us to compare the benefits or not, of the new environment for analyzing with multivariate data.

4. **Does 3D provide advantages over 2D, when exploring low-dimensional projections, for understanding structure in high-dimensional data?**
   Gracia et al. (2016) compares analyst tasks between 2- and 3-D scatterplots. They find modest accuracy and error improvements for distance perception and outlier identification in favor of 3D at the cost of a relatively small increase in task time. Nelson, Cook, and Cruz-Neira (1998) similarly compare a 2D projection and its 3D manipulation space. They find a slight advantage in the sphere test and a large advantage in the cluster test in favor of the 3D manipulation space. A human subjects experiment will be conducted to assess the sensitivity analysis techniques in a 2D vs 3D environment.

1.6 Methodology

This research is interdisciplinary; touring was developed by statisticians to explore physics data. Modern advances in hardware from information technology allow for 3D rending in higher quality and immersion than previously possible.

The research corresponding with RO #1 entails *algorithm design* following and further clarifying the work done in Cook and Buja (1997). In the application of the manual tour, we clarified the creation of the rotation matrix. The key to the matrix is to specify the 2 axes of rotation for the manipulation space and apply Rodrigues' rotation formula (Rodrigues 1840). In the application, attention was given to both pre-compiled tour and human-in-the-loop UCS. We provide an open-source version of the manual tour in the R package, `spinifex`, which has since been published on CRAN. This forms the foundation for future work in the remaining objectives.

For RO #2 is a controlled *experimental study* to explore the efficacy of interactive UCS compared with the benchmarks factors of PCA and the grand tour. This is designed as a within-participant study where each participant performs all factors. the study is balanced by assigning participants into one of 3 groups where the factor order is controlled by a Latin square while simulation order remains the same. The details are discussed in finer detail in section (#sec:expStudy), below.

The research for RO #3 involves *algorithm design* extending the current 2D manual tour into a 3D project. For a 2D projection, the axes basis is rotated through a 3D 'manipulation space'. In a 3D projection, such a space requires 4 dimensions. Theoretically, after the addition of a new angle of rotation, the rotation matrix must be extended to accommodate a new dimension and angle parameter. This also means that analysts have another parameter to define, further increasing their already sizable input-volume.

2 Progress since confirmation

During the candidature confirmation review (27 March 2019) we discussed exploratory data analysis, visualization of high dimensional spaces, covered the literature for tours and 3D rendering for information perception. We concluded with a process for a manual tour that allows for user-controlled steering. The appending the document was a mostly complete R package and respective paper providing an open-source application as well as clarifying the rotation matrix that was outlined in Cook and Buja (1997).

2.1 Publication

The paper has since been accepted in the R Journal and it currently undergoing editorial review. This will be published in the first issue of 2020 and available at journal.r-project.org.

2.2 Software

The R package, `spinifex (v0.1.0)` (Spyrison and Cook 2019), has been approved and hosted for public use on the Comprehensive R Archival Network, CRAN (cran.r-project.org/web/packages/spinifex/)(https://cran.r-project.org/web/packages/spinifex/index.html).

New functionality has been added to the development branch, specifically around the interactive use. A user application is developed with `shiny`(Chang et al. 2018). It allows users to explore their data without the need for coding familiarity. It features interactive or precompiled manual touring. It also contains a gallery for flagging bases which can then further be reviewed or saved as .gif and .png files.

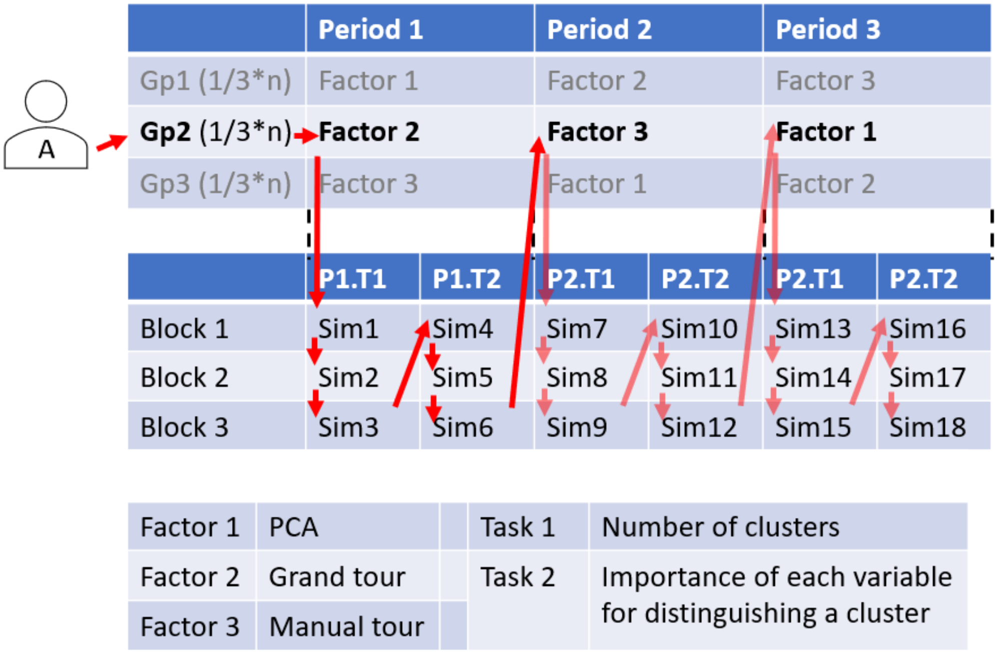
2.3 Experimental study

The prominent appeal of the manual tour is that it allows users to control the contributions of individual variables. In theory, this should enable the finer exploration of features of interest. The hypothesis we will study is *Does the finer control afforded by the manual tour improve the ability of the analyst to understand the importance of variables contributing to the structure?*

We list out an abbreviated summary of the experimental design in the subsections below. A more detailed draft of the working paper is included as an appendix.

2.3.1 Groups

Each participant will be randomly split into one of three even groups. The group controls the order of the factors that the participant was evaluated in for a latin square of the 3 factors. For instance, the order of the first group was PCA, grand, manual. Group level only impacts the order the factors are displayed while task, block, and simulation order will remain the same.



	Period 1	Period 2	Period 3
Gp1 (1/3*n)	Factor 1	Factor 2	Factor 3
Gp2 (1/3*n)	**Factor 2**	**Factor 3**	**Factor 1**
Gp3 (1/3*n)	Factor 3	Factor 1	Factor 2

	P1.T1	P1.T2	P2.T1	P2.T2	P2.T1	P2.T2
Block 1	Sim1	Sim4	Sim7	Sim10	Sim13	Sim16
Block 2	Sim2	Sim5	Sim8	Sim11	Sim14	Sim17
Block 3	Sim3	Sim6	Sim9	Sim12	Sim15	Sim18

Factor 1	PCA		Task 1	Number of clusters
Factor 2	Grand tour		Task 2	Importance of each variable for distinguishing a cluster
Factor 3	Manual tour			

Figure 3: Example case. Person 'A' is assigned to group 2, where they will use factor 2 (grand tour) for the first period. They perform 3 block difficulties of task 1 on simulations of increasing difficulty. Then 3 block difficulties of task 2 on unique simulations sampled from the same distributions of increasing difficulty. After this, they proceed to period 2, where they are use factor 3 (manual tour) to perform 3 block difficulties of each task. Lastly, in the third period, they use factor 1 (PCA) to perform the tasks.

2.3.2 Factors

We will explore performance across three factors. The first factor is Principal Component Analysis (PCA). The second factor is an animated walk of interpolation frames between target bases, called a *grand* tour. The third factor allows for the manual control of the individual variable's contribution to the projection, performing a *manual* tour.

All factors are shown as a scatterplot. The basis axes projection was also illustrated to the left of the plot. They are shown in a unit circle and show the magnitude and direction each variable contributes to the projection.

The user interface was kept the same whenever possible, but the control inputs do change slightly to accommodate the differences between factors. The inputs for PCA select a pair of principal components to display on the x- and y-axes. The manual tour had the same axes selection, with the addition of a drop-down bar and slider control selecting the manipulation variable and magnitude. The grand tour comes precompiled as an animation of a 15 second showing 90 frames at 6 frames per second. The user can control the location or play/pause the animation at will.

2.3.3 Tasks

Within each factor, participants will perform 2 tasks. The first task asks participants to identify the number of clusters present in unsupervised data. This task served as a standard for assessing the general aptitude for this sort of high dimensional analysis as it was simpler. In application, linear discriminant analysis (Fisher 1936) or penalized discriminant analysis (Hastie, Buja, and Tibshirani 1995) are better suited for classifying such unsupervised data.

The second task is focused on the hypothesis of the study, it asks participants to identify any/all variables that were very important and somewhat important for distinguishing a given cluster from the others. For instance, which variables are very- and somewhat- important for distinguishing clusters 'A' and 'B'.

2.3.4 Block difficulty

Participants will be randomly assigned to 1 of 3 even groups. Each group has a different factor order containing all factors. Both tasks will be performed in the same order. Each task will have 3 repetitions performed on new simulations that were drawn from 3 parameterizations in increasing difficulty. Each participant will go through the simulations in the same order, while their factor order will vary. Fixing block difficulty order while varying factors should mitigate potential learning bias.

2.3.5 Pilot study results

Tasks 1 and task 2 aggreageted to take levl:

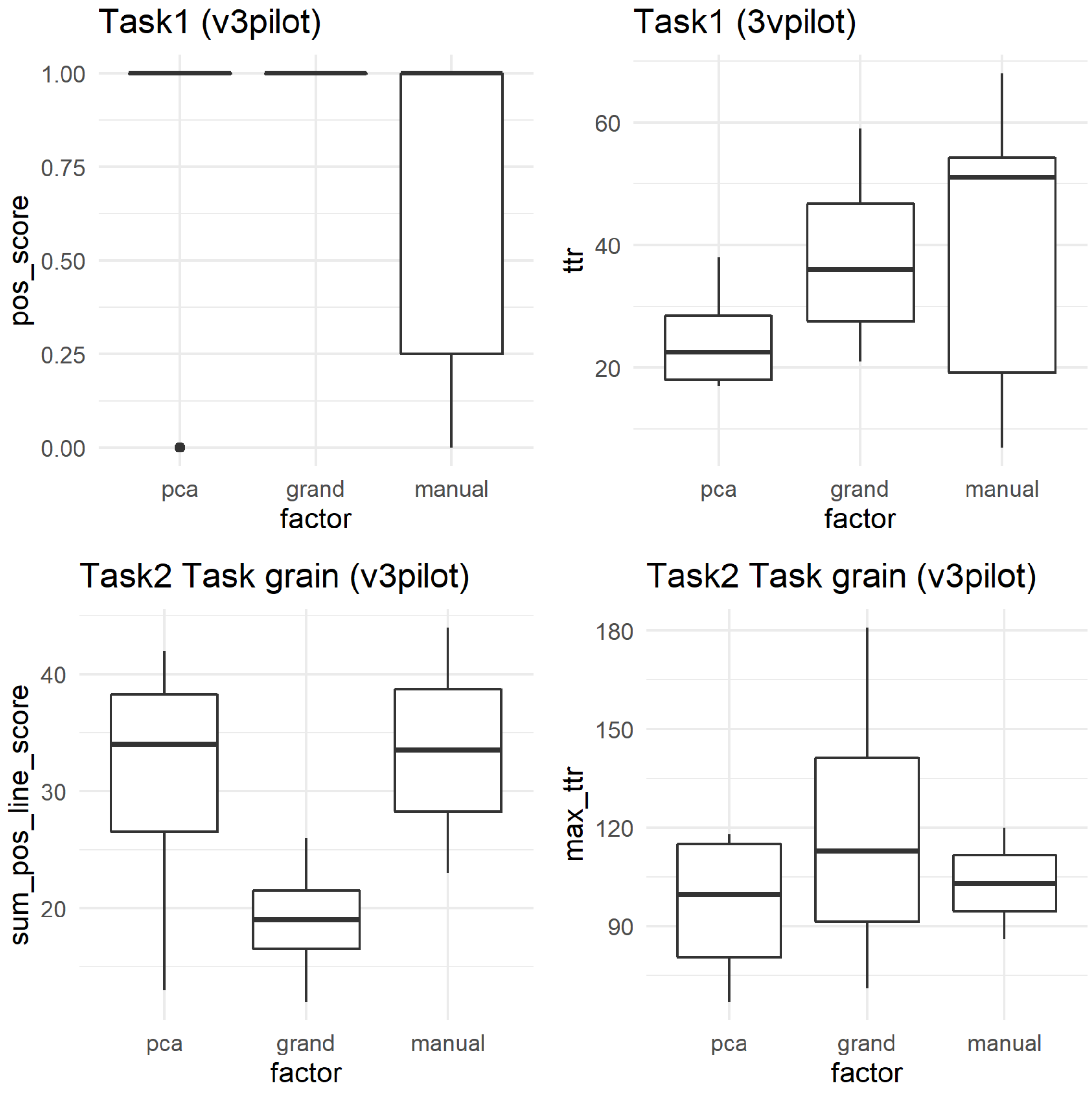


Figure 4: Proposed research timeline.

3 Proposed thesis structure

3.1 Thesis structure

- Introduction – 60%
- Literature review – 80%
- Manual tour and user-controlled steering – 90%
- Experimental study – 60%
- The extension of the manual tour to 3D – 5%
- Conclusion and future plans – 0%

Proposed research timeline

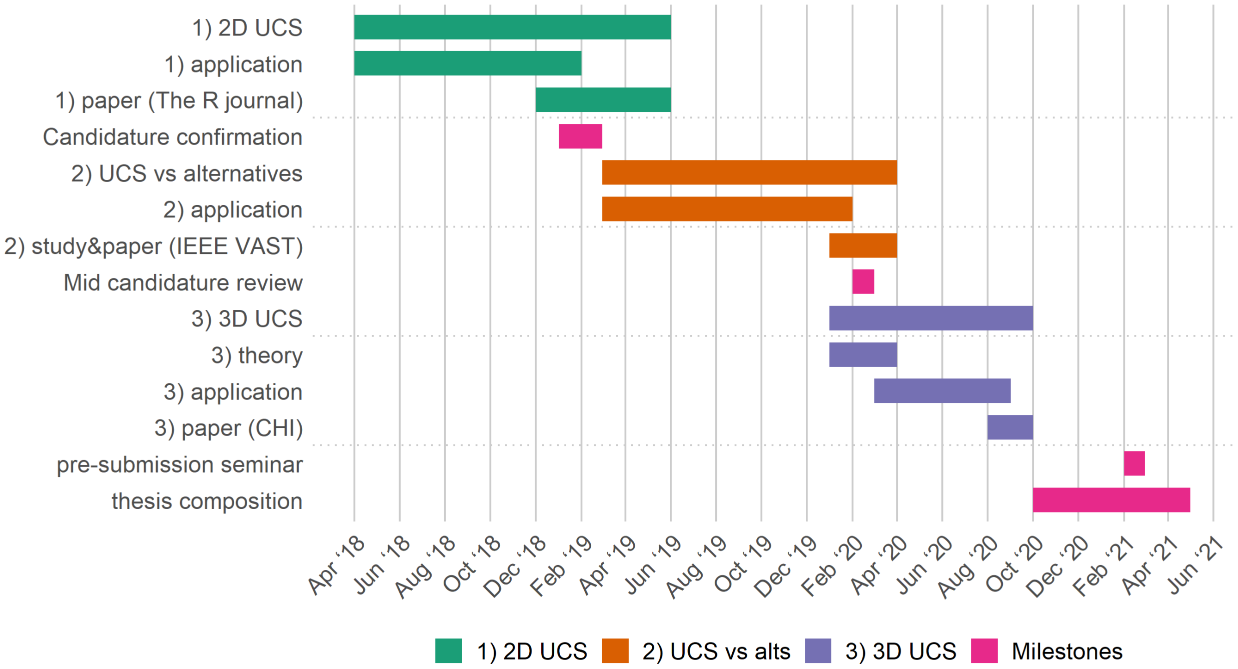


Figure 5: Proposed research timeline.

3.2 Program requirements

- WES Academic record
  - FIT5144: 2019 S1+2, **In progress**, extended to the pre-submission seminar with the unit coordinator for the usual 2 opportunities to complete.

        - Hours: 147>120 hours **Tracked**, missing the following requirments (12 hr total)
        - *Needed:* CYR 2 (A & B) – 2x 3hr
        - *Needed:* Faculty of IT Workshop 1 and 3 on Ethical Research and Publishing – 2x 3hr
    - FIT5113: 2018 S2, **Exemption**
    - FIT6021: 2018 S2, **Completed** with distinction

- myDevelopment - IT: Monash Doctoral Program - Compulsory Module
    - Monash graduate research student induction: **Completed**
    - Research Integrity - Choose the Option most relevant: **Completed**
    - Faculty Induction: **Completed**

4 Potential issues for panel to consider

4.1 Funding for human subjects

- Beverage voucher: $6 x 25 people (est) = $150

4.2 Support for conference travel

Conferences:

CHI 2021: May 8-13, 2021 Yokohama, Japan

submission: Thursday Sep. 10, 2020

https://chi2021.acm.org/

IEEE VAST - VAST 2020: 25-30 October 2020 Salt Lake City, Utah, USA

submission: Saturday, March 21, 2020

http://ieeevis.org/year/2020/info/call-participation/vast-paper-types

5 Acknowledgements

This report was created in `R` (R Core Team 2019) using `rmarkdown` (Xie, Allaire, and Grolemund 2018).

For version control, transparency, and reproducibility, the source files are made available found at github.com/nspyrison/mid_candidature.

References

Anscombe, F. J. 1973. “Graphs in Statistical Analysis.” *The American Statistician* 27 (1): 17–21. https://doi.org/10.2307/2682899.

Asimov, Daniel. 1985. “The Grand Tour: A Tool for Viewing Multidimensional Data.” *SIAM Journal on Scientific and Statistical Computing* 6 (1): 128–43. https://doi.org/https://doi.org/10.1137/0906011.

Buja, Andreas, and Daniel Asimov. 1986. “Grand Tour Methods: An Outline.” In *Proceedings of the Seventeenth Symposium on the Interface of Computer Sciences and Statistics on Computer Science and Statistics*, 63–67. New York, NY, USA: Elsevier North-Holland, Inc. http://dl.acm.org/citation.cfm?id=26036.26046.

Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. “Graphical Methods for Data Analysis.”

Chang, Winston, Joe Cheng, J. J. Allaire, Yihui Xie, and Jonathan McPherson. 2018. *Shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny.

Cook, Dianne, and Andreas Buja. 1997. “Manual Controls for High-Dimensional Data Projections.” *Journal of Computational and Graphical Statistics* 6 (4): 464–80. https://doi.org/10.2307/1390747.

Fisher, Ronald A. 1936. “The Use of Multiple Measurements in Taxonomic Problems.” *Annals of Eugenics* 7 (2): 179–88. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Gabriel, Karl Ruben. 1971. “The Biplot Graphic Display of Matrices with Application to Principal Component Analysis.” *Biometrika* 58 (3): 453–67.

Gracia, Antonio, Santiago González, Víctor Robles, Ernestina Menasalvas, and Tatiana von Landesberger. 2016. “New Insights into the Suitability of the Third Dimension for Visualizing Multivariate/Multidimensional Data: A Study Based on Loss of Quality Quantification.” *Information Visualization* 15 (1): 3–30. https://doi.org/10.1177/1473871614556393.

Hastie, Trevor, Andreas Buja, and Robert Tibshirani. 1995. “Penalized Discriminant Analysis.” *The Annals of Statistics*, 73–102.

Karwowski, Waldemar. 2006. *International Encyclopedia of Ergonomics and Human Factors, -3 Volume Set*. CRC Press.

Matejka, Justin, and George Fitzmaurice. 2017. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 1290–4. Denver, Colorado, USA: ACM Press. https://doi.org/10.1145/3025453.3025912.

Nelson, Laura, Dianne Cook, and Carolina Cruz-Neira. 1998. "XGobi Vs the C2: Results of an Experiment Comparing Data Visualization in a 3-d Immer- Sive Virtual Reality Environment with a 2-d Workstation Display." *Computational Statistics* 14 (1): 39–52.

Pearson, Karl. 1901. "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11): 559–72.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rodrigues, Olinde. 1840. *Des Lois Géométriques Qui Régissent Les Déplacements d'un Système Solide Dans L'espace: Et de La Variation Des Cordonnées Provenant de Ces Déplacements Considérés Indépendamment Des Causes Qui Peuvent Les Produire.*

Spyrison, Nicholas S., and Dianne Cook. 2019. *Spinifex: Manual Tours, Manual Control of Dynamic Projections of Numeric Multivariate Data* (version 0.1.0.9000). https://github.com/nspyrison/spinifex/.

Tukey, John W. 1977. *Exploratory Data Analysis*. Vol. 32. Pearson.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.